Modeling Frequency and Severity of Claims with the Generalized Cluster-Weighted Model

Nik Počuča

McMaster University

Tatjana Miljkovic, Petar Jevtić, and Paul McNicholas

October 1, 2018

Overview

Introduction to Risk

Sub-grouping of insurance policies based on risk classification is a standard practice in insurance. The heterogenous nature of insurance data allows for explorations of many different techniques for sub-grouping risk. As a result, there is a growing number of papers in the area of mixture modeling of univariate and multivariate insurance data to account for heterogeneity of risk.

Examples in Insurance

Automotive

Drivers of various levels of competency are mixed in with large groups rates and are often difficult to track within a cohort.

Health/Life

The variance among people's lifestyles tend to dictate their life expectancy as well as healthcare coverage. Again how do you define a "lifestyle" in a quantitative sense?

Maritime

Maritime Surveillance Radar data is often used to price maritime insurance which have had success being modelled as a mixture of distributions.

Cluster Weighted Models

Let $(\boldsymbol{X}', Y)'$ be the pair of a vector of covariates $\boldsymbol{X}$ and a response variable Y. Assume this set is defined on some sample space Ω that takes values in an appropriate Euclidian subspace. Furthermore, assume that there exists G partitions of Ω, denoted as $\Omega_1, \ldots, \Omega_G$.

Gershenfeld (1997) characterized the cluster-weighted models as a finite mixture of GLMs hence, the joint distribution $f(\boldsymbol{x}, y)$ of $(\boldsymbol{X}', Y)'$ is expressed as follows

$$f(\boldsymbol{x}, y) = \sum_{j=1}^{G} \tau_j q(y|\boldsymbol{x}; \Omega_j) p(\boldsymbol{x}; \Omega_j). \tag{1}$$

Extending CWM

(Ingrassia, Punzo et. al. 2015) proposed a flexible family of mixture models for fitting the joint distribution of a random vector $(\boldsymbol{X}', Y)'$ by splitting the covariates into continuous and discrete as $\boldsymbol{X} = (\boldsymbol{V}', \boldsymbol{W}')'$.

$$
\begin{aligned}
f(\boldsymbol{x}, y; \boldsymbol{\Phi}) &= \sum_{j=1}^{G} \tau_j q(y|\boldsymbol{x}; \boldsymbol{\vartheta}_j) p(\boldsymbol{x}; \boldsymbol{\theta}_j) \\
&= \sum_{j=1}^{G} \tau_j q(y|\boldsymbol{x}; \boldsymbol{\vartheta}_j) p(\boldsymbol{v}; \boldsymbol{\theta}_j^{\star}) p(\boldsymbol{w}; \boldsymbol{\theta}_j^{\star\star})
\end{aligned}
$$

Generalizing CWM

We proceed to extend CWM by splitting the continuous covariates further as $\boldsymbol{V} := (\boldsymbol{U}^{'}, \boldsymbol{T}^{'})^{'}$, where $\boldsymbol{U}$ is a set of non-Gaussian covariates, and $\boldsymbol{T}$ a set of Gaussian covariates. Thus CWM is now recovered as

$$f(\boldsymbol{x}, y; \boldsymbol{\Phi}) = \sum_{j=1}^{G} \tau_j q(y|\boldsymbol{x}; \boldsymbol{\vartheta}_j) p(\boldsymbol{t}; \boldsymbol{\theta}_j^{\star}) p(\boldsymbol{w}; \boldsymbol{\theta}_j^{\star\star}) p(\boldsymbol{u}; \boldsymbol{\theta}_j^{\star\star\star})$$

Non-Gaussian Covariate

With a log-normal assumption for $p(\boldsymbol{u};\boldsymbol{\theta}_j^{\star\star\star})$ we have that $\boldsymbol{u}$ is defined on $\mathbb{R}_+^p, \quad p \in \mathcal{N}$ with parameter vector $\boldsymbol{\theta}_j^{\star\star\star}$ having probability density function as

$$p\left(\boldsymbol{u};\boldsymbol{\theta}_j^{\star\star\star} := (\boldsymbol{\mu}_j^{\star\star\star},\boldsymbol{\Sigma}_j^{\star\star\star})\right)$$

$$= \frac{1}{(\prod_{i=1}^p u_i)|\boldsymbol{\Sigma}_j^{\star\star\star}|(2\pi)^{\frac{p}{2}}} \exp\left[-\frac{1}{2}(\ln \boldsymbol{u} - \boldsymbol{\mu}_j^{\star\star\star})' \boldsymbol{\Sigma}_j^{\star\star\star -1}(\ln \boldsymbol{u} - \boldsymbol{\mu}_j^{\star\star\star})\right].$$

- Extreme Weather Events
- Population Density

Zero - Inflated Poisson

Made famous by Lambert (1992), the zero -inflated Poisson model accounts for the presence of excess zeros in data.

$$f(\boldsymbol{x}, y; \Phi) = \sum_{j=1}^{G} \tau_j \left[q(y = 0|\boldsymbol{x}; \boldsymbol{\vartheta}_j) + q(y > 0|\boldsymbol{x}; \boldsymbol{\vartheta}_j) \right] p(\boldsymbol{t}; \boldsymbol{\theta}_j^{\star}) p(\boldsymbol{w}; \boldsymbol{\theta}_j^{\star\star}) p(\boldsymbol{u}; \boldsymbol{\theta}_j^{\star\star\star}).$$

Zero - Inflated Poisson

$$q(y = 0|\boldsymbol{x};\boldsymbol{\vartheta}_j) = \psi_j + (1-\psi_j)e^{-\lambda_j},$$

$$q(y > 0|\boldsymbol{x};\boldsymbol{\vartheta}_j) = (1-\psi_j)e^{-\lambda_j}\frac{(\lambda_j)^y}{y!}.$$

$$\psi_j = \frac{e^{\tilde{\boldsymbol{x}}\bar{\boldsymbol{\beta}}_j'}}{1+e^{\tilde{\boldsymbol{x}}\bar{\boldsymbol{\beta}}_j'}} \qquad \lambda_j = e^{\tilde{\boldsymbol{x}}\boldsymbol{\beta}_j'}.$$

Bernoulli-Poisson Partitioning Method

$$\Omega^B = \bigcup_{l=1}^{G} \Omega_l^B \qquad f^B(\boldsymbol{x}, y; \Phi) = \sum_{l=1}^{G} \tau_l q^B(y|\boldsymbol{x}; \bar{\boldsymbol{\beta}}_l) p(\boldsymbol{t}; \boldsymbol{\theta}_l^{\star}) p(\boldsymbol{w}; \boldsymbol{\theta}_l^{\star\star}) p(\boldsymbol{u}; \boldsymbol{\theta}_l^{\star\star\star}).$$

$$\psi_l = \frac{e^{\tilde{\boldsymbol{x}}\bar{\boldsymbol{\beta}}_l'}}{1 + e^{\tilde{\boldsymbol{x}}\bar{\boldsymbol{\beta}}_l'}} \qquad q^B(y|\boldsymbol{x}; \bar{\boldsymbol{\beta}}_l) = \begin{cases} \psi_l, & y = 0 \\ 1 - \psi_l, & y > 0 \end{cases}$$

$$\Omega^P = \bigcup_{j=1}^{M} \Omega_j^P \qquad f^P(\boldsymbol{x}, y; \Phi) = \sum_{j=1}^{M} \tau_j q^P(y|\boldsymbol{x}; \boldsymbol{\beta}_j) p(\boldsymbol{t}; \boldsymbol{\theta}_j^{\star}) p(\boldsymbol{w}; \boldsymbol{\theta}_j^{\star\star}) p(\boldsymbol{u}; \boldsymbol{\theta}_j^{\star\star\star}).$$

$$\lambda_j = e^{\tilde{\boldsymbol{x}}\boldsymbol{\beta}_j'}, \qquad q^P(y|\boldsymbol{x}; \lambda_j) = e^{-\lambda_j} \frac{\lambda_j^y}{y!}.$$

Bernoulli-Poisson Partitioning Method

$$\Omega = \Omega^Z = \bigcup_{\substack{l \in \{1,\ldots,G\} \\ j \in \{1,\ldots,M\}}} \Omega^Z_{l,j} := \bigcup_{\substack{l \in \{1,\ldots,G\} \\ j \in \{1,\ldots,M\}}} \Omega^B_l \cap \Omega^P_j =: \bigcup_{k \in \{1,\ldots,K \le M \times G\}} \Omega^Z_k,$$

$$\begin{aligned} q^Z_k(y|\boldsymbol{x}; \bar{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_k) &:= q^B(y|\boldsymbol{x}; \bar{\boldsymbol{\beta}}_k) + (1 - q^B(y|\boldsymbol{x}; \bar{\boldsymbol{\beta}}_k))q^P(y|\boldsymbol{x}; \boldsymbol{\beta}_k) \\ &= q(y = 0|\boldsymbol{x}; \boldsymbol{\vartheta}_k) + q(y > 0|\boldsymbol{x}; \boldsymbol{\vartheta}_k), \quad k \in \{1, \ldots, K\}. \end{aligned}$$

EM Algorithm for CWM (Ingrassia et al, 2016)

E-Step

$$\pi_{ij}{}^{(s)} = E[Z_{ij}|(\boldsymbol{x_i}, y_i); \boldsymbol{\Phi}^{(s)}]$$
$$= \frac{\tau_j{}^{(s)} q(y_i|x_i; \boldsymbol{\beta}_j^{(s)}, \lambda_j^{(s)}) p(t_i; \boldsymbol{\mu}_j^{\star(s)}, \boldsymbol{\Sigma}_j^{\star(s)}) p(w_i; \boldsymbol{\gamma}_j^{(s)}) p(u_i; \boldsymbol{\mu}_j^{\star\star\star(s)}, \boldsymbol{\Sigma}_j^{\star\star\star(s)})}{f(\boldsymbol{x}_i, y_i; \boldsymbol{\Phi}^{(s)})}.$$

M-Step

$$\hat{\tau}_j^{(s+1)} = \frac{1}{n}\sum_{i=1}^{n} \pi_{ij}^{(s)}, \qquad \hat{\boldsymbol{\mu}}_j^{\star(s+1)} = \frac{1}{\sum_{i=1}^{n} \pi_{ij}^{(s)}} \sum_{i=1}^{n} \pi_{ij}^{(s)} \boldsymbol{t}_i, \qquad \hat{\boldsymbol{\gamma}}_{jr}^{(s+1)} = \frac{\sum_{i=1}^{n} \pi_{ij}^{(s)} \omega_i^{rs}}{\sum_{i=1}^{n} \pi_{ij}^{(s)}},$$

$$\widehat{\boldsymbol{\Sigma}}_j^{\star(s+1)} = \frac{1}{\sum_{i=1}^{n} \pi_{ij}^{(s)}} \sum_{i=1}^{n} \pi_{ij}^{(s)} (\boldsymbol{t}_i - \hat{\boldsymbol{\mu}}_j^{(s+1)})(\boldsymbol{t}_i - \hat{\boldsymbol{\mu}}_j^{(s+1)})',$$

M-Step for Log-normal

$$\hat{\boldsymbol{\mu}}_j^{\star\star\star(s+1)} = \frac{1}{\sum_{i=1}^{n} \pi_{ij}^{(s)}} \sum_{i=1}^{n} \pi_{ij}^{(s)} \ln \boldsymbol{u}_i,$$

$$\widehat{\boldsymbol{\Sigma}}_j^{\star\star\star(s+1)} = \frac{1}{\sum_{i=1}^{n} \pi_{ij}^{(s)}} \sum_{i=1}^{n} \pi_{ij}^{(s)} (\ln \boldsymbol{u}_i - \hat{\boldsymbol{\mu}}_j^{\star\star\star(s+1)})(\ln \boldsymbol{u}_i - \hat{\boldsymbol{\mu}}_j^{\star\star\star(s+1)})'.$$

EM Algorithm for Zero-Inflated (Lambert, 1992)

E - Step

$$o_{ik}^{(s)} = \begin{cases} \left[1 + \exp\left(- \tilde{\boldsymbol{x}}_i \bar{\boldsymbol{\beta}}_k{}'^{(s)} - e^{\tilde{\boldsymbol{x}}_i \boldsymbol{\beta}_k'^{(s)}} \right)\right]^{-1}, & y_i = 0 \\ 0 \quad , & y_i > 0. \end{cases}$$

M - Step

$$l_c(\lambda_k; y, \boldsymbol{x}, \boldsymbol{o}_k^{(s)}) = \sum_{i=1}^{n} (1 - o_{ik}^{(s)})(y_i \tilde{\boldsymbol{x}}_i \boldsymbol{\beta}_k' - e^{\tilde{\boldsymbol{x}}_i \boldsymbol{\beta}_k'}). \tag{2}$$

$$l_c(\psi_k; y, \boldsymbol{x}, \boldsymbol{o}_k^{(s)}) = \sum_{i=1}^{n} \left(o_{ik}^{(s)} \tilde{\boldsymbol{x}}_i \bar{\boldsymbol{\beta}}_k{}' - \log\left(1 + e^{\tilde{\boldsymbol{x}}_i \bar{\boldsymbol{\beta}}_k{}'}\right)\right), \tag{3}$$

Comparison of Models

How do we know which model is the best, the zero-inflated or standard Poisson? (Wilson, 2016) demostrates the misuse of the Vuong non-nested t-test (Vuong, 1984). Wilson instead defines a replacement in the form of a LR test.

$$H_0: \quad \psi_k = 0 \qquad vs. \qquad H_a: \quad \psi_k \neq 0.$$

The test statistic φ is defined as

$$\varphi = -2\left[l(\tilde{\lambda_k}; y, \boldsymbol{x}) - l(\lambda_k, \psi_k; y, \boldsymbol{x}) \right]. \tag{4}$$

Application - French Motor Policy

A collection of insurance policy information pertaining to motorists in all 24 regions of France. The dataset is loaded from the CASDatasets package (Dutang, 2014).

Attribute	Description
Policy ID	Unique identifier of the policy holder
Claim Nb	Number of claims during exposure period (0,1,2,3,4)
Exposure	The exposure of policy in years (0–1.5)
Power	Power level of car ordered categorical (12 levels)
Car Age	Car age in years
Driver Age	Age of a legal driver
Brand	Car brands (7 types)
Gas	Diesel or Regular
Region	Regions in France (10 classifications)
Density	Number of inhabitants per km^2
Loss Amount	Portion of claim the insurance policy pays

Modelling Severity

$$LossAmount = Density + CarAge + DriverAge + Region + Power + Gas + \epsilon,$$

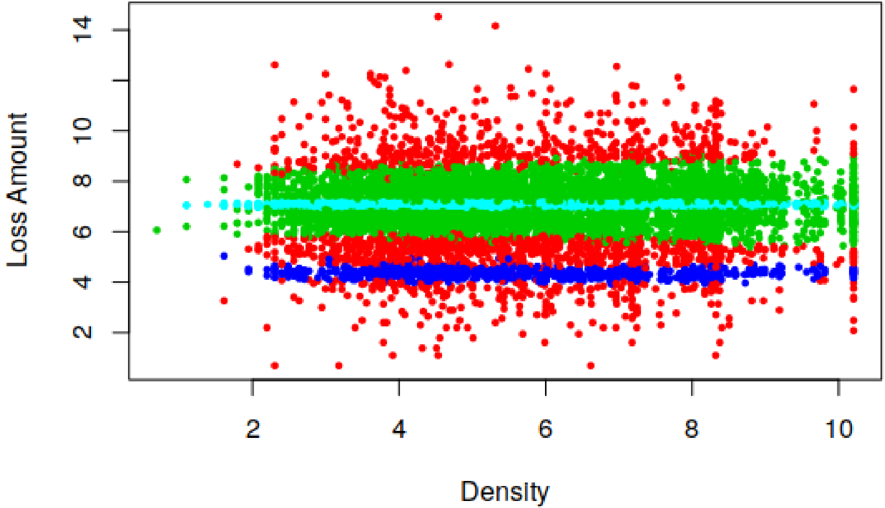
$$\epsilon \sim \mathcal{N}(0, \sigma^2).$$

The canonnical log-link is used for the GLM. The *CarAge* is modelled as a categorical variable with five categories: $[0, 1)$, $[1, 5)$, $[5, 10)$, $[10, 15)$, and $15+$. Additionally, *DriverAge* is modelled as a categorical variable with five categories: $[18, 23)$, $[23, 27)$, $[27, 43)$, $[43, 75)$, and $75+$. *Power* is modelled into three categories as in (Charpentier ,2014).

Comparison of GCWM to CWM

Model	k	AIC	BIC
CWM	1	352, 470	352, 661
	2	314, 560	314, 949
	3	301, 223	301, 812
	4	287, 020	287, 808
	5	**284, 283**	**285, 268**
GCWM	1	111, 129	111, 320
	2	90, 039	90, 428
	3	89, 476	90, 065
	4	**88, 781**	**89, 568**
	5	88, 731	89, 717

Modelling Severity



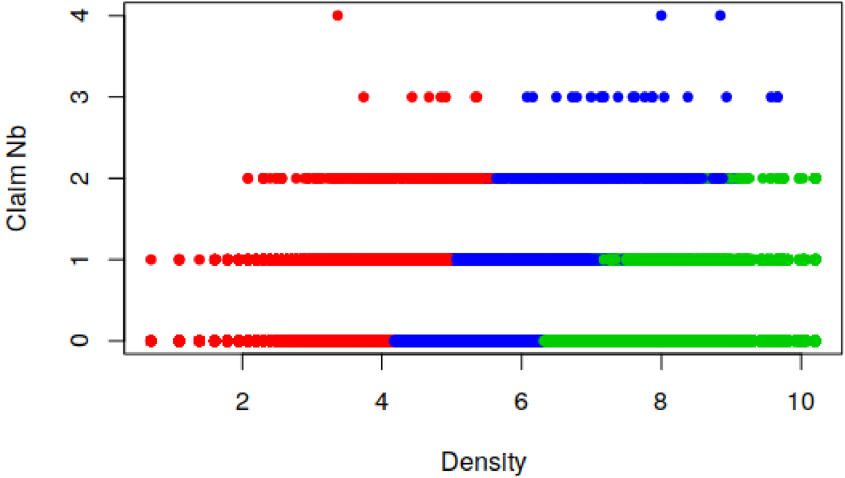
Volatility Clusters

Volatility Level - (Cluster)	Minimum	Mean	Maximum	σ
V1 - (3)	51	79	154	**13**
V2 - (4)	1,039	1,109	1,324	**52**
V3 - (2)	221	1,687	8,841	**1,284**
V4 - (1)	2	9,717	2,036,833	**64,835**

Modelling Frequency

$$ClaimNb = Density + Exposure + Power \quad | \quad Exposure + CarAge \tag{5}$$

Frequency Plot



Density Clusters

Cluster	Color	Minimum	Mean	Maximum	σ
1	Red	0.69	3.38	5.60	0.60
2	Green	6.29	7.86	10.20	1.03
3	Blue	4.13	5.23	9.66	0.65

Conclusions

- GCWM allows for modelling of heterogeneous risk within a set of insurance policies.
- Extension of CWM to GCWM shows improved AIC and BIC.
- Zero-inflated models can also be estimated within the GCWM paradigm.

References

Christophe Dutang (2014)
CAS Datasets

Neil Gershenfeld
Nonlinear Inference and Cluster-Weighted Modeling

Ingrassia, Punzo, et al.
The Generalized Linear Mixed Cluster-Weighted Model
Journal of Classification

Diane Lambert (1992)
Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing
Technometrics

NCDC Storm Events (2018)
NCDC Storm Events

Golden Oak Research Group (2017)
US Household Income

The End